维吾尔文情感分类特征建设研究 *

热西旦木 • 吐尔洪太 [1,2]，吾守尔 • 斯拉木 [1]

(1. 新疆大学 信息科学与工程学院, 乌鲁木齐 830046; 2. 伊犁师范学院 电子与信息工程学院, 新疆 伊宁 835000)

摘　要：由于目前缺乏维吾尔文情感分类特征表示方面的系统性研究，以传统 n-gram 特征为基础，按不同规模从维吾尔文情感标注语料库中提取了新特征及其组合特征，基于支持向量机（SVM）分类器对维吾尔文情感语料库进行了正负情感分类。实验结果表明，所提取的基本特征中 unigram 特征的分类效率最佳；unigram 特征与词组特征的组合可以进一步提高分类效率，其最佳分类效果比 unigram 特征的分类效果提高了 1.78%。首次在统一标注数据集上对不同特征的分类性能进行了综合评价，研究成果可以为今后的维吾尔文情感分类研究提供指导。

关键词：情感分类；特征建设；组合特征；维吾尔文

中图分类号：TP　　doi: 10.3969/j.issn.1001-3695.2018.04.0378

Research on feature construction of Uyghur text sentiment classification

Raxida Turhuntay[1, 2], Wushour Slamu[1]

(1. *College of Information Science & Engineering Xinjiang University*, *Urumqi* 830046, *China*; 2. *College of Electronic & Information Engineering*, *Yili Normal University*, *Yili Xinjiang* 835000, *China*)

Abstract: Due to the lack of systematic research on the feature expression of Uyghur text sentiment classification, this paper uses the traditional n-gram features as the basis to extract new features and combined features from Uyghur sentiment corpora on different scales, classified the corpora as positive and negative with support vector machine (SVM) classifier. Results indicated that, in the Uyghur text sentiment classification, the unigram features in the basic features have the best classification efficiency. The combination of unigram features and phrase features can further improve the classification efficiency. The best performance of the combined features, the classification accuracy is 1.78% higher than that of unigram. This paper first to make a comprehensive evaluation of the classification performance of different features on a unified data set. The research results can be applied as a reference for future Uyghur sentiment classification research.

Key words: sentiment classification; feature construction; combined features; Uyghur

0　引言

文本情感分类在本质上归属于文本分类问题[1]。自康奈尔大学的 Pang 等人[2]将机器学习技术应用于文本情感分类之后，基于机器学习的情感分类技术已获得了广泛的关注和快速的发展。基于机器学习的分类方法经历了浅层学习（传统学习）[3,4]和深度学习[5,6]两次发展浪潮。

基于传统机器学习方法的情感分类研究已经取得了较为丰硕的成果，运用该方法的众多研究工作从特征工程（feature engineering）的范式出发，对情感分类中的特征表示进行了较为深入的研究。在此过程中研究者们较系统地研究了各种不同类型的特征，如 unigram、bigram 等常用的词袋、语法、语义、否定以及组合特征等。

近年来，深度学习方法在文本情感分类研究领域也获得了日益广泛的运用。深度学习模型将从大量无标注语料中自动学习词向量，并将其作为基本特征，从而克服传统方法依靠人工设计特征的不足，能够降低人力和时间成本的消耗。但在将通过深度学习模型训练出的词向量作为情感分类过程的输入特征时，存在一个不容忽视的问题：根据词汇上下文构建词向量时，由于未考虑情感信息，可能发生基于上下文相似而情感极性相反的词汇训练出相似词向量的现象，可能降低情感分类的效率和质量。为解决该问题，研究者们将情感词向量与传统的人工设计特征相结合，以此改善深度学习模型的性能。

相比于汉文、英文等语言的情感分类研究，维吾尔文文本情感分类研究仍处于起步阶段。维吾尔文是形态丰富的黏着性语言，其形态结构远比中文和英文复杂。因此在对维吾尔文文

收稿日期：2018-04-26；**修回日期：**2018-06-25　　**基金项目：**国家“973”重点基础研究计划基金资助项目（2014CB340506）；国家自然科学基金资助项目（61363063）

作者简介：热西旦木 • 吐尔洪太（1980-），女（维吾尔族），新疆伊犁人，讲师，博士研究生，主要研究方向为文本情感分析（raxida522@163.com）；吾守尔·斯拉木（1942-），男（维吾尔族），新疆伊犁人，中国工程院院士，教授，博导，主要研究方向为自然语言处理.

本进行情感分类的过程中，不仅要考虑技术通用性问题，还需要考虑维吾尔文语言的自身特点。目前维吾尔文文本情感分类研究处于初探阶段，有关维吾尔文文本情感分类特征表示方面的研究尚缺乏系统性，因此维吾尔文文本情感分类的大部分研究工作需要从头做起。

本文从自建的维吾尔文情感标注语料库[7]中提取了unigram、bigram、trigram、情感词汇、词性特征、bi-tagged、generalized bi-tagged 等不同的基本特征，并通过 MI（mutual information,互信息）特征选择方法从中提取了最优特征，进而通过组合处理形成了 unigram 与 bigram 的组合特征、unigram 与 bi-tagged 的组合特征以及 unigram 与 generalized bi-tagged 的组合特征。进而在本文情感标注数据集上，对不同特征在维吾尔文情感分类过程中的性能进行了评价与对比。

本文既提取了传统的 n-gram 特征，也提取了体现词汇之间语义关系的多词特征，并初次较系统地在统一标注数据集上对不同特征的性能进行了评价。该工作不仅可以为后续维吾尔文文本情感分类研究工作提供指导，而且还可以为哈萨克语、柯尔克孜语等亲属语言的文本情感分类提供借鉴。

1　相关工作

基于传统机器学习的情感分类方法以人工标注的倾向性文本作为训练集，从中提取情感特征，而后基于机器学习的方法构造情感分类器，再通过训练好的分类器对新文档进行分类。该方法的分类效率很大程度上依赖于对情感特征的质量。国内外已有大量文献比较系统地研究了不同特征对情感倾向标注的影响。

Habernal 等人[3]在自建的 Czech Social Media 情感语料库和电影、产品评论语料库上进行了情感倾向性分类实验，验证了几种预处理方法对情感分类效率的影响，并提取了 n-gram、character n-gram、词性、表情符号等几种基本特征以及在其基础上形成的组合特征，进而基于支持向量机（SVM）和最大熵（MaxEnt）机器学习分类器对数据集进行了情感分类。Rehab 等人[4]研究了词干提取，特征组合和 n-gram 模型等对分类结果的影响并运用 SVM、朴素贝叶斯（NB）和 K-近邻（KNN）等三种分类器在两种数据集上进行了情感分类实验，均获得了较好的分类结果。在汉文情感分类方面，李泽魁等人[8]基于中文微博语料对词、词组、数值和句法特征进行了对比研究，并提出了基于词典规则的情感评分这一新特征，进而通过大量实验与分析，得出了可靠的特征组合。

随着深度学习方法在图像处理和语音识别等方面的成功应用，近期越来越多的研究者将该方法应用于情感分类任务中。国外 Kim[5]采用卷积神经网络（CNN）实现情感分析和问题分类，获得了较好的分类效果。国内梁军等人[6]利用递归自编码模型来主动学习任务的相关特征，避免了人工特征选择，经对比实验证明该模型能够提升情感分类准确率。部分研究者考虑到基于深度学习的词向量特征中情感信息的缺失，将传统人工设计特征与深度学习特征相结合以提高分类效率。孙超红[9]将 unigram、词性（POS）、情感词典等浅层特征与用 Word2vec 训练得到的词向量特征进行融合，基于用 LSTM 改进的 RNN 模型，对微博文本进行情感极性分类。徐莹莹[10]将词向量与传统人工特征相结合，构建了有监督排序模型预测情感强度，该工作在 2016 年 SemEvel（国际标准语义评测）竞赛英文短语情感强度预测任务中获得了第一名的好成绩。

在维吾尔文文本情感分类方面，田生伟等人[11]选取 unigram、bigram、trigram 等特征，采用文档频率、卡方检验、信息增益等特征选择方法，基于朴素贝叶斯、支持向量机、最大熵等分类算法进行了相关研究。热依莱木·帕尔哈提等人[12]基于自建的小规模语料库，提取了区分性单词并对语料进行了两类分类。阿不都萨拉木·达吾提等人[13]将从文献[12]中提取的区分性单词与情感词典相结合进行情感分类，获得更佳的分类效果。李敏等人[14]基于栈式自编码神经网络研究维吾尔文文本情感分类，得到了比传统机器学习算法更高的准确率，其中宏观准确率达到 90.5%。李冬白等人[15]通过 word2vec 得到语料库中每个单词的向量表示，再将词向量与词性特征线性结合，利用栈式自编码算法实现了从大规模无标注隐式情感文本中自动学习特征，并通过 softmax 分类器完成了维吾尔文文本中的隐式情感的自动分类。王树恒等人[16]结合维吾尔语言特征及词汇间的情感特征，实现了基于 word embedding 和双向 LSTM 深度学习的维吾尔文情感分类模型，其实验结果好于 RNN、CNN 和 SVM 等分类器的分类结果。

2　本文实验数据集及其预处理

2.1　维吾尔文评论语料库

该语料库[7]由采集自几个主要的维吾尔文网站的用户对不同主题的评论信息构成。由于评论中包含了丰富的情感信息，所以满足情感语料库所需数据的条件。语料库将每一条评论的情感倾向标注为正面、负面或中性。该语料库总共标注了 15 814 条评论，其中 10 368 条标注为正面、4 515 条标注为负面、931 条标注为中性。语料库具体信息如表 1 所示。

表 1　评论语料库的三种倾向分布表

网站名	中性	正面	负面
Alkuyi	315	878	657
TianShan	407	3 428	741
Putbal	209	6 062	3 117
总共	931	10 368	4 515

由于本文情感分类研究的范围只限于正面和负面两种倾向，所以本文从标注语料中选择了 4 515 条正面评论和 4 515 条负面评论作为实验语料。

2.2　语料库的预处理

维吾尔文具有非常丰富的形态变化和庞大的词汇量，虽然维吾尔文中词干和词缀的数量有限，但是理论上可以组合而成无限多的词汇，其中，绝大多数词汇在语料库中仅仅出现一次

[17]。由此导致在维吾尔文自然语言处理工作中出现特征空间维数极多，以及随之而来的严重的数据相对稀疏问题。因此需要对实验数据进行一些预处理。

2.2.1 维文词法分析器

维吾尔文词法分析器是由新疆大学多语种重点实验室研究开发的预处理工具，其实现了句子边界识别、词干提取、词性标注等多种标注。该工具用统计与规则相结合的方法识别句子[18]。词干提取工作中，将每个词被描述为一个树状结构，用根节点表示词干，孩子节点表示词缀，边表示词干与词缀之间的约束关系。在词干提取过程中充分考虑了维吾尔语在形态变化过程中发生的音变现象[19]。词性标注实现了如表 2 所示的 15 个一级标注规范[20]。

表 2　词性标注集

序号	词性	标注符号	序号	词性	标注符号
1	名词	N	9	叹词	E
2	形容词	A	10	动词	V
3	数词	M	11	拉丁词	LW
4	量词	Q	12	语气词	T
5	副词	D	13	后位词	R
6	介词	P	14	词缀	X
7	模拟词	I	15	标点符号	Y
8	练词	C			

一条维吾尔文句子通过维文词法分析器进行处理后的结果如下：

ئۇ [T=P][S=ئۇ] بولسا [T=V][S=بول] بىر [T=M][S=بىر] ئىچى [T=N][S=ئىچى] تار [T=A][S=تار] ئادەم [T=N][S=ئادەم] <EOS>

(他是一个**小心眼**的人)

其中：“T=”表示词性；“S=”表示单词的词根；<EOS> 是句子结束标记。

2.2.2 预处理过程

1）分词标注　为了得到单词特征，首先要对文本进行分词处理。维吾尔文是一种拼音文字，词语之间以空格和标点符号来分隔。因此分词对于维吾尔文而言不是一个技术问题，可利用空格和标点符号等对维吾尔文文本进行分词。

2)词干提取　在维吾尔文中词干是表达词汇语义的主体部分，而形态后缀是表达语法信息和时态信息的部分[21]。为了减少特征空间维数，避免维数灾难，需要对词汇进行词干提取。完成词干提取之后既能保留原词的基本语义，也能有效降低特征空间维数。例如，“مەكتەپ”（学校）一词通过连接不同词缀可以形成拼写形式略有不同而主干意义相同的词汇。

مەكتەپتە	在学校
مەكتەپتىن	从学校
مەكتەپنىڭ	学校的
مەكتەپكە	去学校
مەكتەپنى	把学校

本文通过维文词法分析器对本文语料进行了词干提取处理。

3）词性标注　词性信息是发掘情感的重要线索。形容词、副词、动词和名词等可以携带重要的情感信息。本文所设计的实验提取了不同词性的词汇作为特征；同时由于本文提出的 bi-tagged 特征是根据文本中的词性前后组合规则来提取信息，所以运用维文词法分析器对本文语料进行了词性标注。

4)停用词去除　维吾尔文情感文本中有一批出现频率较高，却无助于情感分类的词汇，如“man,u,bir,silar”等。如果将这些词汇作为文本特征，则会增加特征空间的维数，降低分类器的性能，因此有必要对这些词汇进行处理。本文通过自建的维吾尔文情感分类停用词表(包含 1 305 个词)，去除了文本中不表达任何情感的停用词。

3　维吾尔文情感特征的选择

3.1　基本特征

1） n-gram 特征　从句子中分别提取 unigram、bigram、trigram 等特征，分别以 F_{uni}、F_{bi}、F_{tri} 表示。

2）情感词汇特征　情感词通常蕴涵着丰富的感情色彩，往往能透露文本所表达的态度和情绪，因此可以将其作为一种重要的特征。本文将作者自建的维吾尔文情感词典[22]中的所有褒义和贬义词作为基本情感特征，并以 F_{dict} 表示。

3)词性特征　词性信息一直被认为是衡量情感表达的一个重要指标。因此，本文选择名词、动词、形容词、副词和叹词等词性作为基本情感特征，以 F_{pos} 表示。

4）bi-tagged 特征　通常传统的基于 n-gram（$n \geq 2$）的特征提取方法会产生高维数的特征向量，高维数不但增大分类难度，而且会延长分类时间。本文受到文献[23]的方法启发，总结了若干词性组合规则，从文本中提取了符合规则的、具有相邻关系和先后顺序的两个单词所构成的短语，并将其命名为 bi-tagged 特征，以 F_{bi-tag} 表示。bi-tagged 特征词性组合规律如表 3 所示。

表 3　bi-tagged 特征词性组合规则

序号	当前词性	下个词性	例子
1	N	A	Kungli parakende 心烦意乱
2	N	V	Erwahi öchmaq　魂飞胆裂
3	A	V	Achiq yötmaq　忍气吞声
4	A	N	Xata qedem　错误的一步
5	A	A	Mihriban ongluq 善良懂事
6	D	A	Bek chirayliq　很漂亮
7	D	V	Ejep yarishiptö　太好看了
8	E	E	Way ësit　真可惜

5） Generalized bi-tagged 特征　虽然 bigram、trigram、bi-tagged 等特征能提取上下文语义信息，但却存在特征过于稀疏、

泛化能力较弱以及特征维数高等不足。 例如，训练语料中有这样两条语句：

① Bügün shundaq tesirlik kënodin birni kurdüm, silerningmu kurüp bëqishinglarni tewsiye qiliman. （我今天看了一部非常感人的电影，也推荐给你们看。）

② Silerge bir tesirlik kitap tonöshtöray. （给你们介绍一本感人的书。）

从训练语料中可以得到“tesirlik këno, shundaq tesirlik, tesirlik kitap”(感人的书、感人的电影、非常感人）等 bigram 和 bi-tagged 特征序列。虽然这些特征很好地表达了情感，但如果“tesirlik hëkaye”(感人的故事）这个特征在测试语料中出现，上述训练语料上训练的分类器无法确定其情感倾向。针对该问题，本文参照文献[24,25]的思路，将本文提取的 bi-tagged 特征前后两个词汇中的一个替换为其所对应的词性标注符号，并命名为 generalized bi-tagged 特征，用 $F_{Gbi-tag}$ 表示。上述例子中，如果将“tesirlik këno, tesirlik hëkaye, tesirlik kitap”等特征中的后词都替换成对应词性符号，将会得到“tesirlik N”(感人的 N）特征，达到不同特征的泛化效果，从而有效保证训练语料中大部分特征的泛化性能。本文对 bi-tagged 特征分别进行了前置词替换（$F_{Gbi-tag_h}$，对前后两个词中的第一个词汇进行词性符号替换）和后置词替换（$F_{Gbi-tag_t}$，对前后两个词中的第二个词汇进行词性符号替换）的方法，通过对比实验确定了最佳的替换方式。

3.2 组合特征

虽然 unigram 特征在情感分类任务中的分类效果总是优于其他特征，但是 unigram 特征的劣势在于不能提取文本中的上下文信息。虽然 bigram、trigram、bi-tagged 等词组特征能提高语义含量，但却降低了特征向量的统计质量，使特征变得更加稀疏，导致机器学习算法难以从中提取可用于分类的统计特性。由于该缺点，采用这些特征获得的情感分类效果逊于采用 unigram 特征的效果[3]。针对该问题，本文对 F_{uni} 和 F_{bi} 特征、F_{uni} 和 F_{bi-tag} 特征、F_{uni} 和 $F_{Gbi-tag-t}$ 特征进行组合，分别形成了 F_{uni-bi}、$F_{uni-bi-tag}$ 和 $F_{uni-Gbi-tag}$ 组合特征。

本文对特征进行组合时，设计了一个组合比例控制参数 α，α=[0.1, 0.2, 0.3, 0.4,…, 0.9]，即按照不同的组合比例对两种特征进行组合，从而确定每一种特征在组合特征中的重要程度。其中 α 是指与 unigram 特征进行组合的词组特征在组合特征总体中所占的比例。以组合特征 $F_{uni-bi-tag}$ 为例，当组合特征总数为 N 时，bi-tagged 特征数 N_{bi-tag}=N*α ，unigram 特征数 N_{uni}=N-N_{bi-tag}，即用 N_{uni} 个 unigram 特征和 N_{bi-tag} 个 bi-tagged 特征进行组合。

4 实验与结果分析

本文从维吾尔文情感语料库中提取不同类型的特征之后，利用 MI 特征选择方法对其进行筛选，利用 tf-idf 特征权重方法判断其区分能力的强弱，随后运用 SVM 机器学习分类器在维吾尔文评论情感语料库上完成正负二元情感分类。实验采用 10 倍交叉验证法，即把数据集分成 10 个子集，在每一轮实验中将其中一个子集作为测试集，其余 9 个子集作为训练集，共执行 10 轮。之后将所得到的结果取平均值作为最终结果。所有实验均用 Python 语言和 Scikit-learn 工具包来完成，实验结果用准确率（Accuracy）来评价。

4.1 基本特征上的分类结果

为了验证本文提取的特征在维吾尔文情感分类过程中的性能，实验从语料集中提取不同特征，并利用 MI 特征选择方法对特征进行排序，从中依次选择排在前 10%到 90%的特征并对不同规模特征对分类器性能的影响进行比较。实验结果描述如表 4 所示。

表 4 基本特征上的分类结果

特征数	F_{uni}	F_{bi}	F_{tri}	F_{dict}	F_{pos}	F_{bi-tag}	$F_{Gbi-tag-h}$	$F_{Gbi-tag-t}$
10%	87.44	**79.17**	60.16	81.49	82.86	74.77	74.24	77.64
20%	89.30	78.65	**63.89**	82.97	84.72	76.41	76.61	79.13
30%	**89.47**	78.69	61.60	83.36	85.28	82.16	78.68	81.85
40%	89.40	78.66	59.59	84.05	85.29	79.31	**81.29**	82.96
50%	89.40	78.34	58.24	84.25	**85.78**	81.96	78.00	84.66
60%	89.26	78.14	57.66	84.17	85.26	**83.99**	76.91	**85.23**
70%	89.28	77.58	54.68	84.30	85.15	79.97	76.98	82.33
80%	89.31	74.87	57.26	84.39	85.13	77.75	77.25	81.85
90%	88.82	72.01	56.01	**84.65**	85.06	77.07	78.87	82.12

由表 4 可知，所有基本特征中 F_{uni} 特征的分类效果最佳。提取前 30%（3324 个特征）的特征时，分类器的分类准确率达到了最高值 89.47，但是随着特征数的增加，其准确率从峰值下降。本文实验中，F_{dict}、F_{pos} 等特征也取得了较理想的结果。例如，从 F_{dict} 中提取 90%（2 316 个特征）的特征时，分类器的分类准确率达到了 86.50；从 F_{pos} 中提取 50%（5 269 个特征）的特征时，分类器的分类准确率达到了 85.78。

基本特征中，F_{bi} 和 F_{tri} 特征对分类结果的影响低于预期。在词组特征中，本文所提出的 bi-tagged 特征与 bigram 特征相比分类效果更佳，其最高分类准确率为 83.99，比 bigram 的最高分类准确率 79.17 高出了 4.82%。对 bi-tagged 特征进行泛化后，其分类效果可以进一步增强，两种 generalized bi-tagged 特征中 $F_{Gbi-tag-t}$ 的分类效率优于 $F_{Gbi-tag-h}$。例如，从 $F_{Gbi-tag-t}$ 中提取前 60%的特征时，分类准确率达到 85.23，比基于同样数目的 F_{bi-tag} 特征的分类效率提高了 1.24%。

4.2 组合特征上的分类结果

为了验证提取的组合特征在维吾尔文情感分类过程中的性能，本文分别运用三种组合特征（F_{uni-bi}、$F_{uni-bi-tag}$、$F_{uni-Gbi-tag}$）在维吾尔文评论情感语料库上进行了情感分类。考虑到 $F_{Gbi-tag-t}$ 的分类效果优于 $F_{Gbi-tag-h}$，对 F_{uni} 和 $F_{Gbi-tag-t}$ 特征进行组合形成组合特征 $F_{uni-Gbi-tag}$。实验过程中，将特征总数从 1 000 逐步增加到 10 000，每次增加 1 000 个特征；将特征数比例控

制参数 α 从 10%逐步增加到 90%。三种组合特征在语料库上分类准确率如表 5~7 所示。限于篇幅，本文仅呈现特征数比例控制参数 α 在相间节点上的结果。

表 5　组合特征 F_{uni-bi} 上的分类结果

α	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
0.1	86.47	88.81	89.24	89.28	89.34	89.26	89.73	89.99	89.40	88.72
0.3	86.31	88.68	89.43	89.60	89.35	89.54	89.92	89.59	89.11	89.13
0.5	85.66	87.89	89.06	89.28	89.48	89.71	89.98	**90.72**	90.35	90.47
0.7	83.82	87.03	87.98	88.69	89.37	89.72	90.67	89.85	90.21	89.75
0.9	79.59	82.31	85.36	86.86	87.89	88.39	88.49	88.69	88.17	88.83

表 6　组合特征 $F_{uni-bi-tag}$ 上的分类结果

α	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
0.1	87.11	89.32	90.00	89.51	89.67	89.71	89.75	89.79	89.81	89.75
0.3	86.42	88.26	89.70	89.99	90.15	90.05	90.20	90.29	90.15	89.98
0.5	85.99	88.19	89.14	90.13	**90.97**	90.18	89.98	90.02	90.02	90.00
0.7	84.29	87.21	88.59	88.82	89.45	89.79	90.05	90.17	90.36	90.20
0.9	80.16	83.88	85.95	86.70	87.26	87.71	88.01	87.98	88.30	88.51

表 7　组合特征 $F_{uni-Gbi-tag}$ 上的分类结果

α	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
0.1	88.94	89.14	89.48	89.43	89.33	89.44	89.13	90.18	90.12	90.17
0.3	86.94	88.96	89.42	89.56	89.67	89.76	90.64	90.75	90.93	89.75
0.5	85.10	88.60	89.29	89.46	89.64	89.83	90.28	**91.25**	90.31	89.84
0.7	84.48	87.22	88.97	89.37	89.36	89.44	89.92	90.48	90.63	90.01
0.9	82.73	85.93	87.05	87.59	87.76	87.92	88.55	89.01	88.14	87.95

在基于组合特征 F_{uni-bi} 的分类实验中（表 5），当特征总数为 8 000，F_{bi} 特征占总特征的 50%时，分类器获得最高分类准确率 90.72，比 F_{uni} 的最高分类准确率提高了 1.25%；

在基于组合特征 $F_{uni-bi-tag}$ 的分类实验中（表 6），当特征总数为 5 000，F_{bi-tag} 特征占总特征的 50%时，分类器获得最高分类准确率 90.97，比 F_{uni} 的最高分类准确率提高了 1.23%，比 F_{uni-bi} 的最高分类准确率提高了 0.25%；

在基于组合特征 $F_{uni-Gbi-tag}$ 的分类实验中（表 7），当特征总数为 8 000，$F_{Gbi-tag-t}$ 特征占总特征的 50%时，分类器获得最高分类准确率 91.25，比 F_{uni}、F_{uni-bi} 和 $F_{uni-Gbi-tag}$ 特征分别提高了 1.78%，0.53%和 0.28%。

实验结果表明，对 unigram 特征与包含上下文语义信息的词组特征进行组合可以有效地客服这些特征各自存在的不足，并可获得比单独使用其中某个特征更佳的分类结果。在基于组合特征的分类实验中，unigram 同与其组合的词组特征在总特征中各占一半时，分类效果最佳。因为当 unigram 特征呈现数据稀疏时，词组特征能够提取一些情感丰富的上下文信息，对 unigram 特征起到补充作用。本文分类实验中 $F_{uni-bi-tag}$ 组合特征的分类效果优于 F_{uni-bi} 组合特征的分类效果。主要原因是 bi-tagged 特征可以删除 bigram 特征中的诸多噪声特征，并能提取结构稳定、语义完整的上下文信息。三种组合特征中，$F_{uni-Gbi-tag}$ 组合特征的分类效率优于前两种组合特征，主要原因是对 bi-tagged 特征进行泛化能进一步提高 bi-tagged 特征的统计特性，可以有效解决其存在的数据稀疏问题，所以 $F_{uni-Gbi-tag}$ 特征的分类效果更佳。

5　结束语

针对目前维吾尔文文本情感分类特征表示相关研究缺乏系统性的问题，本文以传统 n-gram 特征为基础，按不同规模从自建的维吾尔文情感语料库中提取了八种基本特征和三种组合特征（既包含了传统的 BOW 特征，又包含了兼顾上下文信息的语义特征）进行实验。实验证明，在基于基本特征的维吾尔文文本情感分类任务中，unigram 特征的分类效果最佳，若对 unigram 特征与考虑上下文语义信息的词组特征进行组合，能够进一步增强分类效果。本文所涉及的三种组合特征中，unigram 与 generalized bi-tagged 的组合特征分类效果最佳，比 unigram 特征的分类效率提高了 1.78%。

本文所涉及的词组特征是基于词性搭配规则提取具有先后顺序和相邻关系的两个词所组成的特征，目前尚无法以包含两个以上单词的语句为单元进行情感分类。今后的工作将着重研究如何通过拓展词组特征长度及利用长距离词汇之间的依赖关系提高情感分类效率。将本文所提取的特征与深度学习模型的词向量特征进行融合，将其作为深度学习模型的输入特征去评价其在基于深度学习模型的情感分类任务中的性能。

参考文献：

[1] Liu Bing. Sentiment analysis and opinion mining [C]// Proc of Synthesis Lectures on Human Language Technologies. [S. l.] : Morgan & Claypool, 2012: 152-153.

[2] Pang Bo, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [C]// Proc of Acl-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. 2002: 79-86.

[3] Habernal I, Steinberger J. Supervised sentiment analysis in czech social media [M]. [S. l.] : Pergamon Press, Inc, 2014.

[4] Duwairi Rehab, EI-Orfali M. A study of the effects of preprocessing strategies on sentiment analysis for arabic text [J]. Journal of Information Science, 2014, 40 (4): 501-513.

[5] Kim Y. Convolutional neural networks for sentence classification [J]. arXiv: 1408. 5882v2, 2014.

[6] 梁军，柴玉梅，原慧斌，等. 基于深度学习的微博情感分析 [J]. 中文信息学报, 2014, 28 (5): 155-161. (Liang Jun, Chai Yumei, Yuan Huibin, *et al.* Deep learning for Chinese micro-blog sentiment analysis [J]. Journal of Chinese Information Processing, 2014, 28 (5): 155-161.)

[7] 伊尔夏提·吐尔贡，吾守尔·斯拉木，热西旦木·吐尔洪太，等. 维吾尔文情感语料库的构建与分析 [J]. 计算机与现代化, 2017 (4): 67-72. (Yierxiati Tuergong, Wushouer Silamu, Rexidan Tuerhongtai, *et al.* Construction and analysis of Uyghur emotional corpus [J]. Jisuanji Yu

Xiandaihua, 2017 (4): 67-72.)

[8] 李泽魁，赵妍妍，秦兵，等. 中文微博情感倾向性分析特征工程 [J]. 山西大学学报：自然科学版, 2014, 37 (4): 570-579.

[9] (Li Zekui, Zhao Yanyan, Qin Bing, *et al.* Feature engineering for Chinese microblog sentiment classification [J]. Journal of Shanxi University: Natural Science Edition, 2014, 37 (4): 570-579.)

[10] 孙超洪. 基于递归神经网络的微博情感分类研究 [D]. 杭州：浙江理工大学, 2017. (Sun Chaohong. Research on micro-blog sentiment classification based on recurrent neural network [D]. Hangzhou: Master's Thesis of Zhejiang Sci-Tech University, 2017.)

[11] 徐莹莹. 基于深度神经网络模型的句子级文本情感分类研究 [D]. 深圳：深圳大学, 2016. (Xu Yingying. Research of sentence-level sentiment classification for text based on deep neural network [D]. Shenzhen: Master's Thesis of Shenzhen University, 2016.)

[12] 田生伟，禹龙，王宇光. 维吾尔语情感分类算法 [J]. 计算机工程与应用, 2011, 47 (36): 147-150. (Tian Shengwei, Yu Long, Wang Yuguang. Research on sentiment classification of Uighur reviews [J]. Computer Engineering and Applications, 2011, 47 (36): 147-150.)

[13] 热依莱木·帕尔哈提，孟祥涛，艾斯卡尔·艾木都拉. 基于区分性关键词模型的维吾尔文本情感分类 [J]. 计算机工程, 2014, 40 (10): 132-136. (Rayila Parhat, Meng Xiangtao, Askar Hamdulla. Uyghur text sentiment classification based on discriminative keyword model [J]. Computer Engineering, 2014, 40 (10): 132-136.)

[14] 阿不都萨拉木·达吾提，于斯音·于苏普，艾斯卡尔·艾木都拉. 类别区分词与情感词典相结合的维吾尔文句子情感分类 [J]. 清华大学学报：自然科学版, 2017, 57 (2): 197-201. (Abdusalam Dawut, Hussein Yusuf, Askar Hamdulla. Emotion recognition from Uyghur sentences based on combinations of class discrimination words and a sentiment dictionary [J]. Journal of Tsinghua University: Natural Science Edition, 2017, 57 (2): 197-201.)

[15] 李敏，禹龙，田生伟，等. 基于深度学习的维吾尔语语句情感倾向分析 [J]. 计算机工程与设计, 2016, 37 (8): 2213-2217. (Li Min, Yu Long, Tian Shengwei, *et al.* Emotional tendency analysis of Uyghur statement based on deep learning [J]. Computer Engineering and Design, 2016, 37 (8): 2213-2217.)

[16] 李冬白，田生伟，禹龙，等. 深度学习的维吾尔语语句隐式情感分类 [J]. 计算机工程与设计, 2016, 37 (9): 2577-2580. (Li Dongbai, Tian Shengwei, Yu Long, *et al.* Deep learning for implicit sentiment classification of Uyghur sentence [J]. Computer Engineering and Design, 2016, 37 (9): 2577-2580.)

[17] 王树恒，吐尔根·依布拉音，卡哈尔江·阿比的热西提，等. 基于BLSTM 的维吾尔语文本情感分析 [J]. 计算机工程与设计, 2017, 38 (10): 2879-2886. (Wang Shuheng, Turgun Ibrahim, Kahaerjiang Aviderexiti, *et al.* Sentiment classification of Uyghur text based on BLSTM [J]. Computer Engineering and Design, 2017, 38 (10): 2879-2886.)

[18] Abdukelimu H, Liu Yang, Chen Xinxiong, *et al.* Learning distributed representations of Uyghur words and morphems [M]. Cham: Springer International Publishing, 2015: 202-211.

[19] 艾山·吾买尔，吐尔根·依步拉音. 统计与规则相结合的维吾尔语句子边界识别 [J]. 计算机工程与应用, 2010, 46 (14): 162-165. (Aishan Wumaier, Tuergen Yibulayin. Sentence boundary detection of Uyghur based on rules and statistics [J]. Computer Engineering and Applications, 2010, 46 (14): 162-165.)

[20] 麦热哈巴·艾力，姜文斌，王志洋，等. 维吾尔语词法分析的有向图模型 [J]. 软件学报, 2012, 23 (12): 94-100. (Maierhaba Aili, Jiang Wenbin, Wang Zhiyang, *et al.* Directed graph model of Uyghur morphological analysis [J]. Journal of Software, 2012, 23 (12): 94-100.)

[21] Maimaiti M, Wumaier A, Abiderexiti K, *et al.* Bidirectional long short-term memory network with a conditional random field layer for Uyghur part-of-speech tagging [J]. Information, 2017, 8 (4): 157.

[22] 力提甫·托乎提. 现代维吾尔语参考语法 [M]. 北京：中国社会科学出版社, 2012. (Litip Tohti. The reference grammar of modern Uyghur language [M]. Beijing: China Social Sciences Press, 2012.)

[23] 热西旦木·吐尔洪太，吾守尔·斯拉木，伊尔夏提·吐尔贡. 词典与机器学习方法相结合的维吾尔语文本情感分析 [J]. 中文信息学报. 2017, 31 (1): 177-183. (Rexidanmu Tuerhongtai, Wushour Silamu, Yierxiati Tuergong. Uyghur text sentiment analysis by combining lexical knowledge with machine learning methods [J]. Journal of Chinese Information Processing, 2017, 31 (1): 177-183.)

[24] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]// Proc of Annual Meeting of the Association for Computational Linguistics. 2002: 417-424.

[25] Gamon M. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis [C]// Proc of International Conference on Computational Linguistics. 2004: 841-847.

[26] Joshi M, Penstein-Rosé C. Generalizing dependency features for opinion mining [C]// Proc of the ACL-IJCNLP 2009 Conference Short Papers. 2009: 313.